Journal of Computer and System Sciences 78 (2012) 1538–1556

Contents lists available at SciVerse ScienceDirect

Journal of Computer and System Sciences

www.elsevier.com/locate/jcss

The K-armed dueling bandits problem

Yisong Yue [a,*], Josef Broder [b], Robert Kleinberg [c], Thorsten Joachims [c]

[a] H. John Heinz III College, Carnegie Mellon University, Pittsburgh, PA 15213, United States
[b] Center for Applied Mathematics, Cornell University, Ithaca, NY 14853, United States
[c] Department of Computer Science, Cornell University, Ithaca, NY 14853, United States

ARTICLE INFO

Article history:
Received 28 March 2010
Received in revised form 31 January 2011
Accepted 22 December 2011
Available online 21 January 2012

Keywords:
Online learning
Multi-armed bandits
Preference elicitation

ABSTRACT

We study a partial-information online-learning problem where actions are restricted to noisy comparisons between pairs of strategies (also known as bandits). In contrast to conventional approaches that require the absolute reward of the chosen strategy to be quantifiable and observable, our setting assumes only that (noisy) binary feedback about the relative reward of two chosen strategies is available. This type of relative feedback is particularly appropriate in applications where absolute rewards have no natural scale or are difficult to measure (e.g., user-perceived quality of a set of retrieval results, taste of food, product attractiveness), but where pairwise comparisons are easy to make. We propose a novel regret formulation in this setting, as well as present an algorithm that achieves information-theoretically optimal regret bounds (up to a constant factor).

1. Introduction

In partial information online learning problems (also known as bandit problems) [27], an algorithm must choose, in each of T consecutive iterations, one of K possible bandits (strategies). For conventional bandit problems, in every iteration, each bandit receives a real-valued payoff in $[0, 1]$, initially unknown to the algorithm. The algorithm then chooses one bandit and receives (and thus observes) the associated payoff. No other payoffs are observed. The goal then is to maximize the total payoff (i.e., the sum of payoffs over all iterations).

The conventional setting assumes that observations perfectly reflect (or are unbiased estimates of) the received payoffs. In many applications, however, such observations may be unavailable or unreliable. Consider, for example, applications in sensory testing or information retrieval, where the payoff is the goodness of taste or the user-perceived quality of a retrieval result. While it is difficult to elicit payoffs on an absolute scale in such applications, one can reliably obtain relative judgments of payoff (i.e. "A tastes better than B", or "ranking A is better than ranking B"). In fact, user behavior can often be modeled as maximizing payoff, so that such relative comparison statements can be derived from observable user behavior. For example, to elicit whether a search-engine user prefers ranking r_1 over r_2 for a given query, Radlinski et al. [26] showed how to present an interleaved ranking of r_1 and r_2 so that clicks indicate which of the two is preferred by the user. This ready availability of pairwise comparison feedback in applications where absolute payoffs are difficult to observe motivates our learning framework.

Given a collection of K bandits (also referred to as arms, actions or strategies), we wish to find a sequence of noisy comparisons that has low regret. We call this the *K-armed Dueling Bandits Problem*, which can also be viewed as a regret-minimization version of the classical problem of finding the maximum element of a set using noisy comparisons [14]. This paper extends results originally published in [28] with a more thorough and refined theoretical analysis.

[*] Corresponding author.
E-mail addresses: yisongyue@cmu.edu (Y. Yue), jbroder@cam.cornell.edu (J. Broder), rdk@cs.cornell.edu (R. Kleinberg), tj@cs.cornell.edu (T. Joachims).

0022-0000/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcss.2011.12.028

A canonical application example of the dueling bandits problem is an intranet-search system that is installed for a new customer. Among K built-in retrieval functions, the search engine needs to select the one that provides the best results on this collection, with pairwise feedback coming from clicks in the interleaved rankings [26]. Since the search engine incurs regret whenever it presents the results from a suboptimal retrieval function, it aims to efficiently identify and eliminate suboptimal retrieval functions in order to maximize user satisfaction. More generally, the dueling bandits problem arises naturally in many applications where a system must adapt interactively to specific user bases, and where pairwise comparisons are easier to elicit than absolute payoffs.

One important issue is formulating an appropriate notion of regret. Since we are concerned with maximizing user utility (or satisfaction), but utility is not directly quantifiable in our pairwise-comparison model, a natural question to ask is whether users, at each iteration, would have preferred another bandit over the ones chosen by our algorithm. This leads directly to our regret formulation (described in Section 3), which measures regret based on the (initially unknown) probability that the best bandit b^* would win a comparison with the chosen bandits at each iteration. One can alternatively view this as the fraction of users who would have preferred b^* over the bandits chosen by our algorithm.

Our solution follows an “explore then exploit” approach, where we will bound expected regret by the regret incurred while running the exploration algorithm. We will present an exploration algorithm in Section 4, which we call Interleaved Filter. Interleaved Filter incurs regret that matches (up to constant factors) the information-theoretic optimum in expectation, and is within a logarithmic factor with high probability. We will prove the matching lower bound in Section 10.

An interesting feature of our Interleaved Filter algorithms is that, unlike previous search algorithms based on noisy comparisons, e.g., [14], the number of experiments devoted to each bandit during the exploration phase is highly non-uniform: of the K bandits, there is a small subset of bandits ($\mathcal{O}(\log K)$ of them) who each participate in $\mathcal{O}(K)$ comparisons, while the remaining bandits only participate in $\mathcal{O}(\log K)$ comparisons in expectation. In Section 10 we provide insight about why existing methods suffer high regret in our setting.

2. Related work

Regret-minimizing algorithms for multi-armed bandit problems and their generalizations have been intensively studied for many years, both in the stochastic [20] and non-stochastic [3] cases. The vast literature on this topic includes algorithms whose regret is within a constant factor of the information-theoretic lower bound in both the stochastic case [2] and the non-stochastic case [1]. Our use of upper confidence bounds in designing algorithms for the dueling bandits problem is prefigured by their use in the multi-armed bandit algorithms that appear in [6,2,20].

Upper confidence bounds are also central to the design of multi-armed bandit problems in the PAC setting [12,24], where the algorithm's objective is to identify an arm that is ε-optimal with probability at least $1-\delta$. Our work adopts a very different feedback model (pairwise comparisons rather than direct observation of payoffs) and a different objective (regret minimization rather than the PAC objective), but there are clear similarities between our proposed algorithms and the Successive Elimination and Median Elimination algorithms developed for the PAC setting in [12]. There are also some clear differences between the algorithms: these are discussed in Section 11.

The difficulty of the dueling bandits problem stems from the fact that the algorithm has no way of directly observing the costs of the actions it chooses. It is an example of a *partial monitoring problem*, a class of regret-minimization problems defined in [9], in which an algorithm (the “forecaster”) chooses actions and then observes feedback signals that depend on the actions chosen by the forecaster and by an unseen opponent (the “environment”). This pair of actions also determines a loss, which is not revealed to the forecaster but is used in defining the forecaster's regret. Under the crucial assumption that the feedback matrix has high enough rank that its row space spans the row space of the loss matrix (which is required in order to allow for a Hannan consistent forecaster) the results of [9] show that there is a forecaster whose regret is bounded by $O(T^{2/3})$ against a non-stochastic (adversarial) environment, and that there exist partial monitoring problems for which this bound cannot be improved. Our dueling bandits problem is a special case of the partial monitoring problem. In particular, our environment is stochastic rather than adversarial, and thus our regret bound exhibits much better (i.e., logarithmic) dependence on T.

Banditized online learning problems based on absolute rewards (of individual actions) have been previously studied in the context of web advertising [25,22]. In that setting, clear explicit feedback is available in the form of (expected) revenue. We study settings where such absolute measures are unavailable or unreliable.

Our work is also closely related to the literature on computing with noisy comparison operations [4,8,14,18], in particular the design of tournaments to identify the maximum element in an ordered set, given access to noisy comparators. All of these papers assume unit cost per comparison, whereas we charge a different cost for each comparison depending on the pair of elements being compared. In the unit-cost-per-comparison model, and assuming that every comparison has ϵ probability of error regardless of the pair of elements being compared, Feige et al. [14] presented sequential and parallel algorithms that achieve the information-theoretically optimal expected cost (up to constant factors) for many basic problems such as sorting, searching, and selecting the maximum. The upper bound for noisy binary search has been improved in a recent paper [8] that achieves the information-theoretic optimum up to a $1+o(1)$ factor. When the probability of error depends on the pair of elements being compared (as in our dueling bandits problem), Adler et al. [4] and Karp and Kleinberg [18] present algorithms that achieve the information-theoretic optimum (up to constant factors) for the problem of selecting the maximum and for binary search, respectively. Our results can be seen as extending this line of work to the setting of regret minimization. It is worth noting that the most efficient algorithms for selecting the maximum in the model of noisy comparisons with unit cost per comparison [4,14] are not suitable in the regret minimization setting considered here, because they devote undue effort to comparing elements that are far from the maximum. This point is discussed further in Section 11.

Yue and Joachims [29] simultaneously studied a continuous version of the dueling bandits problem, where bandits (e.g., retrieval functions) are characterized using a compact and convex parameter space. For that setting, they proposed a gradient descent algorithm which achieves sublinear regret (with respect to the time horizon). In many applications, it may be infeasible or undesirable to interactively explore such a large space of bandits. For instance, in intranet search one might reasonably "cover" the space of plausible retrieval functions with a small number of hand-crafted retrieval functions. In such cases, selecting the best of K well-engineered solutions would be much more efficient than searching a possibly huge space of real-valued parameters.

Learning based on pairwise comparisons is well studied in the (off-line) supervised learning setting called learning to rank. Typically, a preference function is first learned using a set of i.i.d. training examples, and subsequent predictions are made to minimize the number of mis-ranked pairs (e.g., [10]). Most prior work assume access to a training set with absolute labels (e.g., of relevance or utility) on individual examples, with pairwise preferences generated using pairs of inputs with labels from different ordinal classes (e.g., [5,7,13,15,17,21]). In the case where there are exactly two label classes, this becomes the so-called bipartite ranking problem [5,7], which is a more general version of learning to optimize ROC-Area [15,17,21].

3. The dueling bandits problem

We propose a new online optimization problem, called the K-armed Dueling Bandits Problem, where the goal is to find the best among K bandits $\mathcal{B} = \{b_1, \dots, b_K\}$. Each iteration comprises a noisy comparison (a duel) between two bandits (possibly the same bandit with itself). We assume that the outcomes of these noisy comparisons are independent random variables and that the probability of b winning a comparison with b' is stationary over time. We write this probability as $P(b > b') = \epsilon(b, b') + 1/2$, where $\epsilon(b, b') \in (-1/2, 1/2)$ is a measure of the distinguishability between b and b'. Note that $\epsilon(b, b') = -\epsilon(b', b)$ and $\epsilon(b, b) = 0$. We assume that there exists a total ordering on $\mathcal{B}$ such that $b \succ b'$ implies $\epsilon(b, b') > 0$. We will also use the notation $\epsilon_{i,j} \equiv \epsilon(b_i, b_j)$.

Let $(b_1^{(t)}, b_2^{(t)})$ be the bandits chosen at iteration t, and let b^* be the overall best bandit. We define **strong regret** based on comparing the chosen bandits with b^*,

$$R_T = \sum_{t=1}^{T} \max\{\epsilon(b^*, b_1^{(t)}), \epsilon(b^*, b_2^{(t)})\}, \tag{1}$$

where T is the time horizon. We also define **weak regret**,

$$\tilde{R}_T = \sum_{t=1}^{T} \min\{\epsilon(b^*, b_1^{(t)}), \epsilon(b^*, b_2^{(t)})\}, \tag{2}$$

which only compares $\hat{b}$ against the better of $b_1^{(t)}$ and $b_2^{(t)}$. One can regard regret as the fraction of users who would have preferred the best bandit over the chosen ones in each iteration.[1] More precisely, it corresponds to the fraction of users who prefer the best bandit to the worse of the pair of bandits chosen, in the case of strong regret, or to the better of the two bandits chosen, in the case of weak regret. Note that our results are immediately applicable to any notion of regret that is a linear combination of weak and strong regret, e.g.

$$\frac{1}{2}\left(\epsilon(b^*, b_1^{(t)}) + \epsilon(b^*, b_2^{(t)})\right).$$

We will present algorithms which achieve identical regret bounds for both notions of regret (up to constant factors) by assuming a property called stochastic triangle inequality, which is described in the next section.

3.1. Modeling assumptions

We impose additional structure to the probabilistic comparisons. First, we assume **strong stochastic transitivity**, which requires that any triplet of bandits $b_i \succ b_j \succ b_k$ satisfies

$$\epsilon_{i,k} \geqslant \max\{\epsilon_{i,j}, \epsilon_{j,k}\}. \tag{3}$$

This assumption places a monotonicity or internal consistency constraint on possible probability values.

[1] In the search setting, users experience an interleaving, or mixing, of results from both retrieval functions to be compared.

Algorithm 1 Explore then exploit solution.

```
Input: $T$, $\mathcal{B} = \{b_1, \ldots, b_K\}$, $EXPLORE$
$(\hat{b}, \hat{T}) \leftarrow EXPLORE(T, \mathcal{B})$
for $t = \hat{T} + 1, \ldots, T$ do
    compare $\hat{b}$ and $\hat{b}$
end for
```

We also assume **stochastic triangle inequality**, which requires any triplet of bandits $b_i \succ b_j \succ b_k$ to satisfy

$$\epsilon_{i,k} \leqslant \epsilon_{i,j} + \epsilon_{j,k}. \tag{4}$$

Stochastic triangle inequality captures the condition that the probability of a bandit winning (or losing) a comparison will exhibit diminishing returns as it becomes increasingly superior (or inferior) to the competing bandit.[2]

We briefly describe two common generative models which satisfy these two assumptions. The first is the logistic or Bradley–Terry model, where each bandit b_i is assigned a positive real value μ_i. Probabilistic comparisons are made using

$$P(b_i > b_j) = \frac{\mu_i}{\mu_i + \mu_j}.$$

The second is a Gaussian model, where each bandit is associated with a random variable X_i that has a Gaussian distribution with mean μ_i and variance 1. Probabilistic comparisons are made using

$$P(b_i > b_j) = P(X_i - X_j > 0),$$

where $X_i - X_j \sim N(\mu_i - \mu_j, 2)$. It is straightforward to check that both models satisfy strong stochastic transitivity and stochastic triangle inequality. We will describe and justify a more general family of probabilistic models in Appendix A.

4. Algorithm and main results

Our solution, which is described in Algorithm 1, follows an "explore then exploit" approach. For a given time horizon T and a set of K bandits $\mathcal{B} = \{b_1, \ldots, b_K\}$, an exploration algorithm (denoted generically as EXPLORE) is used to find the best bandit b^*. EXPLORE returns both its solution $\hat{b}$ as well as the total number of iterations $\hat{T}$ for which it ran (it is possible that $\hat{T} > T$). Should $\hat{T} < T$, we enter an exploit phase by repeatedly choosing $(b_1^{(t)}, b_2^{(t)}) = (\hat{b}, \hat{b})$, which incurs no additional regret assuming EXPLORE correctly found the best bandit ($\hat{b} = b^*$). In the case where $\hat{T} > T$, then the regret incurred from running EXPLORE still bounds our regret formulations (which only measure regret up to T), so our analysis in this section will still hold.[3]

Our exploration algorithm, which we call Interleaved Filter (IF), is described in Algorithm 2. We will show that IF correctly return the best bandit with probability at least $1 - 1/T$. Correspondingly, a suboptimal bandit is returned with probability at most $1/T$, in which case we assume maximal regret $\mathcal{O}(T)$. We can thus bound the expected regret by

$$\begin{aligned}\mathbf{E}[R_T] &\leqslant \left(1 - \frac{1}{T}\right)\mathbf{E}\big[R_T^{IF}\big] + \frac{1}{T}\mathcal{O}(T) \\ &= \mathcal{O}\big(\mathbf{E}\big[R_T^{IF}\big] + 1\big) \end{aligned} \tag{5}$$

where R_T^{IF} denotes the regret incurred from running Interleaved Filter. Thus the regret bound depends entirely on the regret incurred by IF. We will show that IF achieves an expected regret bound which matches the information-theoretic lower bound (up to constant factors) presented in Section 10, and also matches with high probability[4] the lower bound up to a log factor.

Interleaved Filter maintains a candidate bandit $\hat{b}$ and simulates simultaneously comparing $\hat{b}$ with all other remaining bandits via round robin scheduling (i.e., interleaving). Any bandit that is empirically inferior to $\hat{b}$ with $1 - \delta$ confidence is removed (we will describe later how to choose δ). When some bandit b' is empirically superior to $\hat{b}$ with $1 - \delta$ confidence, then $\hat{b}$ is removed and b' becomes the new candidate $\hat{b} \leftarrow b'$. Afterwards, all empirically inferior bandits (even if lacking $1 - \delta$ confidence) are removed (called pruning – see lines 16–18 in Algorithm 2). This process repeats until only one bandit remains. Assuming IF has not made any mistakes, then it will return the best bandit $\hat{b} = b^*$.

Terminology. We use the term "**bandit**" (or "**action**" or "**arm**") to refer to an element of the strategy set $\mathcal{B}$. Interleaved Filter makes a "**mistake**" if it draws a false conclusion regarding a pair of bandits. A mistake occurs when an inferior bandit

[2] Our analysis also applies for a relaxed version where $\epsilon_{i,k} \leqslant \lambda(\epsilon_{i,j} + \epsilon_{j,k})$ for finite $\lambda > 1$.

[3] In practice, we can terminate EXPLORE after it has run for T time steps, in which case the incurred regret is strictly less than running EXPLORE to completion.

[4] We will say that a sequence of random variables $X_{T,K}$ is $\mathcal{O}(f(T,K))$ *with high probability* if for any sufficiently large d, there exists m depending only on d such that $P(X_{T,K} \geqslant mf(T,K)) \leqslant (TK)^{-d}$ for all sufficiently large T, K.

Algorithm 2 Interleaved Filter (IF).

```
 1: Input: $T$, $\mathcal{B} = \{b_1, \dots, b_K\}$
 2: $\delta \leftarrow 1/(TK^2)$
 3: Choose $\hat{b} \in \mathcal{B}$ randomly
 4: $W \leftarrow \{b_1, \dots, b_K\} \setminus \{\hat{b}\}$
 5: $\forall b \in W$, maintain estimate $\hat{P}_{\hat{b},b}$ of $P(\hat{b} > b)$ according to (6)
 6: $\forall b \in W$, maintain $1-\delta$ confidence interval $\hat{C}_{\hat{b},b}$ of $\hat{P}_{\hat{b},b}$ according to (7), (8)
 7: while $W \neq \emptyset$ do
 8:   for $b \in W$ do
 9:     compare $\hat{b}$ and $b$
10:     update $\hat{P}_{\hat{b},b}$, $\hat{C}_{\hat{b},b}$
11:   end for
12:   while $\exists b \in W$ s.t. $(\hat{P}_{\hat{b},b} > 1/2 \wedge 1/2 \notin \hat{C}_{\hat{b},b})$ do
13:     $W \leftarrow W \setminus \{b\}$   //$\hat{b}$ declared winner against $b$
14:   end while
15:   if $\exists b' \in W$ s.t. $(\hat{P}_{\hat{b},b'} < 1/2 \wedge 1/2 \notin \hat{C}_{\hat{b},b'})$ then
16:     while $\exists b \in W$ s.t. $\hat{P}_{\hat{b},b} > 1/2$ do
17:       $W \leftarrow W \setminus \{b\}$   //pruning
18:     end while
19:     $\hat{b} \leftarrow b'$,  $W \leftarrow W \setminus \{b'\}$   //$b'$ declared winner against $\hat{b}$ (new round)
20:     $\forall b \in W$, reset $\hat{P}_{\hat{b},b}$ and $\hat{C}_{\hat{b},b}$
21:   end if
22: end while
23: $\hat{T} \leftarrow$ Total Comparisons Made
24: return $(\hat{b}, \hat{T})$
```

is determined with $1-\delta$ confidence to be the superior one. We call the elimination step in lines 16–18 of Algorithm 2 "**pruning**". We define a "**match**" to be all the comparisons Interleaved Filter makes between two bandits, and a "**round**" to be all the matches played by one candidate $\hat{b}$. We always refer to $\log x$ as the natural log, $\ln x$, whenever the distinction is necessary.

In our analysis, we assume without loss of generality that the bandits in $\mathcal{B}$ are indexed in preferential order $b_1 \succ \dots \succ b_K$. We first state our main result.

Theorem 1. *Running Algorithm 1 with $\mathcal{B} = \{b_1, \dots, b_K\}$, time horizon T $(T \geqslant K)$, and IF incurs expected regret (both weak and strong) bounded by*

$$\mathbf{E}[R_T] = \mathcal{O}\big(\mathbf{E}\big[R_T^{IF}\big]\big) = \mathcal{O}\left(\frac{K}{\epsilon_{1,2}} \log T\right).$$

Theorem 1 follows immediately from combining Lemma 1 and Lemma 3 below. Note that $\epsilon_{1,2} = P(b_1 > b_2) - 1/2$ is the distinguishability between the two best bandits. Due to strong stochastic transitivity, $\epsilon_{1,2}$ lower bounds the distinguishability between the best bandit and any other bandit.

Analysis approach. Our analysis follows three phases. We first bound the regret incurred for any match. Then we show that the probability of IF making a mistake is at most $1/T$. We finally bound the matches played by IF to arrive at our final regret bounds. The behavior of Interleaved Filter can be summarized by the following three lemmas, which we will prove in Section 7, Section 8, and Section 9, respectively.

Lemma 1. *The probability that IF makes a mistake resulting in the elimination of the best bandit b_1 is at most $1/T$.*

Lemma 2. *Assuming IF is mistake-free, then, with high probability,*

$$R_T^{IF} = \mathcal{O}\left(\frac{K \log K}{\epsilon_{1,2}} \log T\right)$$

for both strong and weak regret.

Lemma 3. *Assuming IF is mistake-free, then*

$$\mathbf{E}\big[R_T^{IF}\big] = \mathcal{O}\left(\frac{K}{\epsilon_{1,2}} \log T\right)$$

for both strong and weak regret.

5. Confidence intervals

In a match between b_i and b_j, Interleaved Filter (IF) maintains a number

$$\hat{P}_{i,j} = \frac{\#\, b_i \text{ wins}}{\#\, \text{comparisons}}, \tag{6}$$

which is the empirical estimate of $P(b_i > b_j)$ after t comparisons.[5] For ease of notation, we drop the subscripts (b_i, b_j), and use $\hat{P}_t$, which emphasizes the dependence on the number of comparisons. IF also maintains a confidence interval

$$\hat{C}_t = (\hat{P}_t - c_t, \hat{P}_t + c_t), \tag{7}$$

where

$$c_t = \sqrt{4\log(1/\delta)/t}. \tag{8}$$

We justify the construction of these confidence intervals in the following lemma.

Lemma 4. *For $\delta = 1/(TK^2)$, the number of comparisons in a match between b_i and b_j is with high probability at most*

$$\mathcal{O}\left(\frac{1}{\epsilon_{i,j}^2}\log(TK)\right).$$

Moreover, the probability that the inferior bandit is declared the winner at some time $t \leqslant T$ is at most δ.

Proof. First we argue that the probability of the inferior bandit being declared the winner is at most δ. Note that by the stopping condition of the match, if IF mistakenly declares the inferior bandit the winner at time t, then we must have $1/2 + \epsilon_{i,j} \notin \hat{C}_t$ (note that $\epsilon_{i,j}$ can be either positive or negative). By the definition of $\hat{C}_t$ and the fact that $\mathbf{E}[\hat{P}_t] = 1/2 + \epsilon_{i,j}$, we have $P(1/2 + \epsilon_{i,j} \notin \hat{C}_t) = P(|\hat{P}_t - \mathbf{E}[\hat{P}_t]| \geqslant c_t)$. It follows from Hoeffding's inequality [16] that the probability of making a mistake at time t is bounded above by

$$P\big(\big|\hat{P}_t - \mathbf{E}[\hat{P}_t]\big| \geqslant c_t\big) \leqslant 2\exp\big(-2tc_t^2\big) = 2\exp\big(-8\log(1/\delta)\big) = 2\delta^8 = \frac{2}{T^8K^{16}}.$$

Now an application of the union bound shows that the probability of making a mistake at any time $t \leqslant T$ is bounded above by

$$P\left(\bigcup_{t=1}^{T}\{1/2 + \epsilon_{i,j} \notin \hat{C}_t\}\right) \leqslant \frac{2T}{T^8K^{16}} \leqslant \frac{1}{TK^2} = \delta,$$

provided that $K \geqslant 2$, which is the desired result.

We now show that the number of comparisons n in a match between b_i and b_j is $\mathcal{O}(\log(TK)/\epsilon_{i,j}^2)$ with high probability. Specifically, we will show that for any $d \geqslant 1$, there exists an m depending only on d such that

$$P\left(n \geqslant \frac{m}{\epsilon_{i,j}^2}\log(TK)\right) \leqslant (TK)^{-d}$$

for all T, K sufficiently large. By the stopping condition of the match, if at any time t we have $\hat{P}_t - c_t > 1/2$, then the match terminates. It follows that for any time t, if $n > t$, then $\hat{P}_t - c_t \leqslant 1/2$, and so

$$P(n > t) \leqslant P(\hat{P}_t - c_t \leqslant 1/2).$$

To bound this probability, assume without loss of generality that $\epsilon_{i,j} > 0$, and note that since $\mathbf{E}[\hat{P}_t] = 1/2 + \epsilon_{i,j}$, we have

$$P(\hat{P}_t - c_t \leqslant 1/2) = P(\hat{P}_t - 1/2 - \epsilon_{i,j} \leqslant c_t - \epsilon_{i,j}) = P\big(\mathbf{E}[\hat{P}_t] - \hat{P}_t \geqslant \epsilon_{i,j} - c_t\big).$$

For any $m \geqslant 8$ and $t \geqslant \lceil 2m\log(TK^2)/\epsilon_{i,j}^2\rceil$, we have $c_t \leqslant \epsilon_{i,j}/2$, and so applying Hoeffding's inequality for this m and t shows

$$P\big(\mathbf{E}[\hat{P}_t] - \hat{P}_t \geqslant \epsilon_{i,j} - c_t\big) \leqslant P\big(\big|\hat{P}_t - \mathbf{E}[\hat{P}_t]\big| \geqslant \epsilon_{i,j}/2\big) \leqslant 2\exp\big(-t\epsilon_{i,j}^2/2\big).$$

Since $t \geqslant 2m\log(TK^2)/\epsilon_{i,j}^2$ by assumption, we have $t\epsilon_{i,j}^2/2 \geqslant m\log(TK^2)$, and so

[5] In other words, $\hat{P}_{i,j}$ is the fraction of these t comparisons in which b_i was the winner.

$$2\exp\left(-t\epsilon_{i,j}^2/2\right) \leqslant 2\exp\left(-m\log\left(TK^2\right)\right) = \frac{2}{T^m K^{2m}} \leqslant (TK)^{-m}$$

for $T \geqslant 1$ and $K \geqslant 2$, which proves the claim. □

6. Regret per match

We now bound the accumulated regret (both weak and strong) of each match.

Lemma 5. *Assuming b_1 has not been removed and $T \geqslant K$, then with high probability the accumulated weak regret and also strong regret from any match is at most*

$$\mathcal{O}\left(\frac{1}{\epsilon_{1,2}}\log T\right).$$

Proof. Suppose the candidate bandit $\hat{b} = b_j$ is playing a match against b_i. Since all matches within a round are played simultaneously, then by Lemma 4, any match played by b_j contains at most

$$\mathcal{O}\left(\frac{1}{\epsilon_{1,j}^2}\log(TK)\right) = \mathcal{O}\left(\frac{1}{\epsilon_{1,2}^2}\log(TK)\right)$$

comparisons, where the inequality follows from strong stochastic transitivity. Note that $\min\{\epsilon_{1,j}, \epsilon_{1,i}\} \leqslant \epsilon_{1,j}$. Then the accumulated weak regret (2) is bounded by

$$\begin{aligned}\epsilon_{1,j}\mathcal{O}\left(\frac{1}{\epsilon_{1,j}^2}\log(TK)\right) &= \mathcal{O}\left(\frac{1}{\epsilon_{1,j}}\log(TK)\right) \\ &= \mathcal{O}\left(\frac{1}{\epsilon_{1,2}}\log(TK)\right) \\ &= \mathcal{O}\left(\frac{1}{\epsilon_{1,2}}\log T\right) \qquad (9)\end{aligned}$$

where (9) holds since $\log(TK) \leqslant \log(T^2) = 2\log T$. We now bound the accumulated strong regret (1) by leveraging stochastic triangle inequality. Each comparison incurs $\max\{\epsilon_{1,j}, \epsilon_{1,i}\}$ regret. We now consider three cases.

Case 1. Suppose $b_i \succ b_j$. Then $\max\{\epsilon_{1,j}, \epsilon_{1,i}\} = \epsilon_{1,j}$, and the accumulated strong regret of the match is bounded by

$$\epsilon_{1,j}\mathcal{O}\left(\frac{1}{\epsilon_{1,j}^2}\log(TK)\right) = \mathcal{O}\left(\frac{1}{\epsilon_{1,2}}\log(TK)\right).$$

Case 2. Suppose $b_j \succ b_i$ and $\epsilon_{j,i} \leqslant \epsilon_{1,j}$. Then

$$\begin{aligned}\max\{\epsilon_{1,j} + \epsilon_{1,i}\} &= \epsilon_{1,i} \\ &\leqslant \epsilon_{1,j} + \epsilon_{j,i} \\ &\leqslant 2\epsilon_{1,j}\end{aligned}$$

and the accumulated strong regret is bounded by

$$\begin{aligned}2\epsilon_{1,j}\mathcal{O}\left(\frac{1}{\epsilon_{1,j}^2}\log(TK)\right) &= \mathcal{O}\left(\frac{1}{\epsilon_{1,j}}\log(TK)\right) \\ &= \mathcal{O}\left(\frac{1}{\epsilon_{1,2}}\log(TK)\right).\end{aligned}$$

Case 3. Suppose $b_j \succ b_i$ and $\epsilon_{j,i} > \epsilon_{1,j}$. Then we can also use Lemma 4 to bound with high probability the number of comparisons by

$$\mathcal{O}\left(\frac{1}{\epsilon_{j,i}^2}\log(TK)\right).$$

The accumulated strong regret is then bounded by

$$\begin{aligned}2\epsilon_{j,i}\mathcal{O}\left(\frac{1}{\epsilon_{j,i}^2}\log(TK)\right) &= \mathcal{O}\left(\frac{1}{\epsilon_{j,i}}\log(TK)\right)\\ &= \mathcal{O}\left(\frac{1}{\epsilon_{1,j}}\log(TK)\right)\\ &= \mathcal{O}\left(\frac{1}{\epsilon_{1,2}}\log(TK)\right).\end{aligned}$$

Like in the analysis for weak regret (9), we finally note that

$$\mathcal{O}\left(\frac{1}{\epsilon_{1,2}}\log(TK)\right) = \mathcal{O}\left(\frac{1}{\epsilon_{1,2}}\log T\right). \qquad \square$$

7. Mistake bound

We now prove Lemma 1, which bounds the probability that IF makes a mistake by $1/T$. It suffices to consider the following two cases: (A) an inferior bandit defeats the candidate, and (B) a superior bandit was removed during the pruning step (lines 16–18 in Algorithm 2). Case (A) is relatively straightforward to prove, and our analysis focuses primarily on Case (B) in this section.

Lemma 6. *For all triples of bandits $b, b', \hat{b}$ such that $b \succ b'$, the probability that IF eliminates b in a pruning step in which b' wins a match against the incumbent bandit $\hat{b}$ (i.e. $\hat{P}_{\hat{b},b'} < 1/2$) while b is found to be empirically inferior to $\hat{b}$ (i.e. $\hat{P}_{\hat{b},b} > 1/2$) is at most δ.*

Proof. Let $X_1, X_2, \ldots$ denote an infinite sequence of i.i.d. Bernoulli random variables with $\mathbf{E}[X_i] = P(\hat{b} \succ b')$, and let $Y_1, Y_2, \ldots$ denote an infinite sequence of i.i.d. Bernoulli random variables with $\mathbf{E}[Y_i] = P(\hat{b} \succ b)$. We couple the outcomes of the comparisons performed by the algorithm to the sequences (X_i), (Y_i) in the obvious way: X_i (resp. Y_i) represents the outcome of the ith comparison between $\hat{b}$ and b' (resp. $\hat{b}$ and b) if the algorithm performs at least i comparisons of that pair of bandits; otherwise X_i (resp. Y_i) does not correspond to any comparison observed by the algorithm.

If b is eliminated by IF in a pruning step at the end of a match consisting of n comparisons between b' and the incumbent $\hat{b}$, then $X_1, \ldots, X_n$ represent the outcomes of the n matches between $\hat{b}$ and b' in that round, and $Y_1, \ldots, Y_n$ represent the outcomes of the n matches between $\hat{b}$ and b in that round. From the definition of confidence intervals in IF we know that $X_1 + \cdots + X_n < n/2 - \sqrt{4n\log(1/\delta)}$, whereas the definition of the pruning step implies that $Y_1 + \cdots + Y_n > n/2$. Thus, if we define $Z_i = Y_i - X_i$ for $i = 1, 2, \ldots,$ then we have

$$Z_1 + \cdots + Z_n > \sqrt{4n\log(1/\delta)}. \tag{10}$$

To complete the proof of the lemma, we will show the probability that there exists an n satisfying (10) is at most δT.

The random variables $(Z_i)_{i=1}^{\infty}$ are i.i.d. and satisfy $|Z_i| \leqslant 1$. Furthermore, our assumption that $b \succ b'$ together with strong stochastic transitivity implies that

$$\mathbf{E}[Z_i] = P(\hat{b} \succ b) - P(\hat{b} \succ b') \leqslant 0.$$

By Hoeffding's inequality, for every n the probability that $\sum_{i=1}^{n} Z_i$ exceeds $\sqrt{4n\log(1/\delta)}$ is at most $\exp(-8n\log(1/\delta)/(4n)) = \delta^2$. Taking the union bound over $n = 1, 2, \ldots, T$, we find that the probability that there exists an n satisfying (10) is at most $\delta^2 T \leqslant \delta$, as claimed. $\square$

Proof of Lemma 1. By Lemma 4, for every i the probability that b_1 is eliminated in a match against b_i is at most δ. A union bound over all i implies that the probability of b_1 being eliminated by directly losing a match to some other bandit is at most $\delta(K-1)$. On the other hand, by Lemma 6, for all i, j the probability that b_1 is eliminated in a pruning step resulting from a match in which b_i defeats b_j is at most δ. A union bound over all i, j implies that the probability of b_1 being eliminated in a pruning step is at most $\delta(K-1)^2$. Summing these two bounds, the probability that IF makes a mistake resulting in the elimination of b_1 is at most $\delta[(K-1) + (K-1)^2] < \delta K^2 = 1/T$. $\square$

8. High probability exploration bound

In this section, we prove Lemma 2, which claims that mistake-free executions of IF satisfy

$$R_T^{IF} = \mathcal{O}\left(\frac{K\log K}{\epsilon_{1,2}}\log T\right)$$

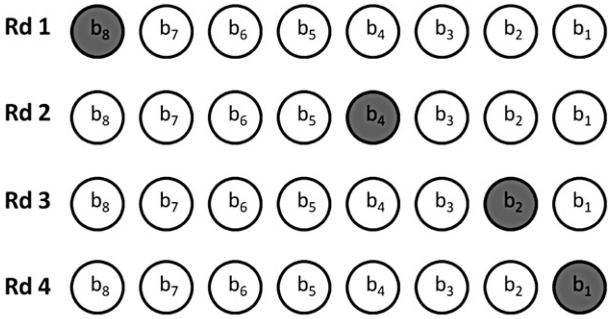

Fig. 1. An illustrative example of a sequence of candidate bandits. The incumbent candidate in each round is shaded in grey.

with high probability. This is within a log factor of the information-theoretic lower bound described in Section 10. The analysis relies on proving that the number of candidate bandits selected by IF (i.e. the number of rounds) is $\mathcal{O}(\log K)$ with high probability.

We can model the sequence of candidate bandits as a random walk. Fig. 1 contains an illustrative example. Let b_j denote the incumbent candidate, and let p_i denote the probability that b_i will be the incumbent in the following round. Due to strong stochastic transitivity, we know that $p_{j-1} \leqslant \cdots \leqslant p_1$. We will see that the "worst case" scenario is the case where all stochastic preferences are equal, i.e. $\epsilon_{j,k} = \epsilon$. In this case, we have $p_{j-1} = \cdots = p_1 = 1/(j-1)$ (assuming no mistakes are made). This setting can be captured in the following Random Walk Model.

Definition 1 *(Random Walk Model).* Define a random walk graph with K nodes labeled $b_1, \ldots, b_K$ (these will correspond to the similarly named bandits). Each node b_j $(j > 1)$ transitions to b_i for $j > i \geqslant 1$ with probability $1/(j-1)$, or in other words b_j transitions to $b_1, \ldots, b_{j-1}$ with uniform probability. The final node b_1 is an absorbing node.

It is clear to see that a path in the Random Walk Model corresponds to a sequence of candidate bandits taken by IF in an instance of the dueling bandits problem where $\epsilon_{1,j} = \epsilon_{2,j} = \cdots = \epsilon_{j-1,j}$ for all $j > 1$ (and no mistakes are made). Thus, the path length of the random walk is exactly to the number of rounds in IF in that case. We will use the Random Walk Model to bound the number of rounds in mistake-free executions of IF.

Proposition 1. *Either IF makes a mistake, or else the number of rounds in the execution of IF is stochastically dominated by the path length of a random walk in the Random Walk Model. In other words, if S and $\tilde{S}$ are random variables corresponding to the number of rounds in IF and the Random Walk Model, respectively, then*

$$\forall x: \quad P(S \geqslant x) \leqslant P(\tilde{S} \geqslant x).$$

As a consequence, any bound on the path length of the random walk also upper bounds the number of rounds in mistake-free executions of IF.

Proposition 1 follows directly from Lemma 14 in Appendix B. This allows us to concentrate our analysis on the (simpler) upper bound setting of the Random Walk Model. We will prove that the random walk in the Random Walk Model requires $\mathcal{O}(\log K)$ steps with high probability. Let X_i $(1 \leqslant i < K)$ be an indicator random variable corresponding to whether a random walk starting at b_K visits b_i in the Random Walk Model. We first analyze the marginal probability of each $P(X_i = 1)$, and also show that $X_1, \ldots, X_{K-1}$ are mutually independent.

Lemma 7. *Let X_i be as defined above with $1 \leqslant i < K$. Then*

$$P(X_i = 1) = \frac{1}{i},$$

and furthermore, for all $W \subseteq \{X_1, \ldots, X_{K-1}\}$, we can write $P(W) \equiv P(\bigwedge_{i \in W} X_i)$ as

$$P(W) = \prod_{X_i \in W} P(X_i), \tag{11}$$

meaning $X_1, \ldots, X_{K-1}$ are mutually independent.

Proof. We can rewrite (11) as

$$P(W) = \prod_{X_i \in W} P(X_i \mid W_i),$$

where $W_i = \{X_j \in W \mid j > i\}$.

We first consider $W = \{X_1, \ldots, X_{K-1}\}$. For the factor on X_i, denote with j the smallest index in W_i with $X_j = 1$ in the condition. Then

$$P(X_i = 1 \mid X_{i+1}, \ldots, X_{K-1}) = P(X_i = 1 \mid X_{i+1} = 0, \ \ldots, \ X_{j-1} = 0, \ X_j = 1) = \frac{1}{i}, \tag{12}$$

since, conditioned on $X_{i+1} = 0, \ldots, \ X_{j-1} = 0, \ X_j = 1$, the random walk must move to one of the first i nodes with uniform probability. Since (12) holds for all $j > i$, this implies $P(X_i = 1) = \frac{1}{i}$. So we can conclude

$$P(X_1, \ldots, X_{K-1}) = \prod_{i=1}^{K-1} P(X_i).$$

Now consider arbitrary W. We use $\sum_{W^c}$ to indicate summing over the joint states of all X_i variables not in W. We can write $P(W)$ as

$$\begin{aligned}
P(W) &= \sum_{W^c} P(X_1, \ldots, X_{K-1}) \\
&= \sum_{W^c} \prod_{i=1}^{K-1} P(X_i) \\
&= \prod_{X_i \in W} P(X_i) \left(\sum_{W^c} \prod_{X_i \in W^c} P(X_i) \right) \\
&= \prod_{X_i \in W} P(X_i).
\end{aligned}$$

This proves mutual independence (11). □

We can express the number of steps taken by a random walk from b_K to b_1 in the Random Walk Model as

$$S_K = 1 + \sum_{i=1}^{K-1} X_i. \tag{13}$$

Lemma 7 implies that

$$E[S_K] = 1 + \sum_{i=1}^{K-1} E[X_i] = 1 + H_{K-1} \approx \log K,$$

where H_i is the harmonic sum. We now show that $S_K = \mathcal{O}(\log K)$ with high probability. We first remark that one can easily prove $P(X_i = 1) \leqslant 1/i$ without using the Random Walk Model, if one defines the probabilities of the random walk of candidates using the actual stochastic preferences $\epsilon_{i,j}$. However, we also require that each X_i be independent of the others in order to prove a high probability bound on S_K.

Lemma 8. *Assuming IF is mistake-free, then it runs for $\mathcal{O}(\log K)$ rounds with high probability.*

Proof. Due to Proposition 1, it suffices to analyze the distribution of path lengths in the Random Walk Model. It thus suffices to show that for any d sufficiently large, there exists an m depending only on d such that

$$\forall K \geqslant 1: \quad P(S_K > m \log K) \leqslant \frac{1}{K^d}, \tag{14}$$

for S_K as defined in (13). From Lemma 7, we know that the random variables $X_1, \ldots, X_{K-1}$ in S_K are mutually independent. Then using the Chernoff bound [23], we know that for any $m > 1$,

$$\begin{aligned}
P\big(S_K > m(1 + H_{K-1})\big) &\leqslant \left(\frac{e^{m-1}}{m^m} \right)^{1+H_{K-1}} \\
&\leqslant \left(\frac{e^{m-1}}{m^m} \right)^{1+\log K} \\
&= (eK)^{m-1-m\log m}
\end{aligned} \tag{15}$$

(15) is true since

$$\log K \leqslant H_{K-1} < \log K + 1$$

for all $K \geqslant 1$. We require this bound to be at most $1/K^d$, or

$$(eK)^{m-1-m\log m} \leqslant K^{-d}.$$

The above inequality is satisfied by $m \geqslant d$ for $d \geqslant e$. The Chernoff bound applies for all $K \geqslant 0$. So for any $d \geqslant e$, we can choose $m = d$ to satisfy (14). □

Corollary 1. *Assuming IF is mistake-free, then it plays $\mathcal{O}(K \log K)$ matches with high probability.*

Proof. The result immediately follows from Lemma 8 by noting that IF plays at most $\mathcal{O}(K)$ matches in each round. □

Proof of Lemma 2. The result immediately follows from combining Lemma 5 and Corollary 1. □

9. Expected Regret Bound

In Section 9, we showed a high probability regret bound that is withing a log factor of the information-theoretic lower bound described in Section 10. We now prove Lemma 3, which claims that mistake-free executions of IF satisfy

$$\mathbf{E}\big[R_T^{IF}\big] = \mathcal{O}\left(\frac{K}{\epsilon_{1,2}} \log T\right),$$

which matches the information-theoretic lower bound up to constant factors.

Lemma 9. *Assuming IF is mistake-free, then it plays $\mathcal{O}(K)$ matches in expectation.*

Proof. Let B_j denote a random variable counting the number of matches played by b_j when it is *not* the incumbent (to avoid double-counting). We can write B_j as

$$B_j = A_j + G_j,$$

where A_j indicates the number of matches played by b_j against b_i for $i > j$ (when the incumbent was inferior to b_j), and G_j indicates the number of matches played by b_j against b_i for $i < j$ (when the incumbent was superior to b_j). We can thus bound the expected number of matches played via

$$\sum_{j=1}^{K-1} \mathbf{E}[B_j] = \sum_{j=1}^{K-1} \mathbf{E}[A_j] + \mathbf{E}[G_j]. \tag{16}$$

By Lemma 7 and leveraging the Random Walk Model defined in Section 8, we can write $\mathbf{E}[A_j]$ as

$$\mathbf{E}[A_j] \leqslant 1 + \sum_{i=j+1}^{K-1} \frac{1}{i} = 1 + H_{K-1} - H_i,$$

where H_i is the harmonic sum.

We now analyze $\mathbf{E}[G_j]$. We assume the worst case that b_j does not lose a match (with $1-\delta$ confidence) to any superior incumbent b_i before the match concludes (b_i is defeated) unless $b_i = b_1$. We can thus bound $\mathbf{E}[G_j]$ using the probability that b_j is pruned at the conclusion of each round. Let $\mathcal{E}_{j,t}$ denote the event that b_j is pruned after the tth round in which the incumbent bandit is superior to b_j, conditioned on not being pruned in the first $t-1$ such rounds. Define $G_{j,t}$ to indicate the number of matches beyond the first $t-1$ played by b_j against a superior incumbent, conditioned on playing at least $t-1$ such matches. We can write $\mathbf{E}[G_{j,t}]$ as

$$\mathbf{E}[G_{j,t}] = 1 + P\big(\mathcal{E}_{j,t}^c\big)\mathbf{E}[G_{j,t+1}],$$

and thus

$$\mathbf{E}[G_j] \leqslant \mathbf{E}[G_{j,1}] \leqslant 1 + P\big(\mathcal{E}_{j,1}^c\big)\mathbf{E}[G_{j,2}]. \tag{17}$$

We know that $P(\mathcal{E}_{j,t}^c) \leqslant 1/2$ for all $j \neq 1$ and t. From Lemma 8, we know that $\mathbf{E}[G_{j,t}] \leqslant \mathcal{O}(K \log K)$ and is thus finite. Hence, we can bound (17) by the infinite geometric series $1 + 1/2 + 1/4 + \cdots = 2$.

We can thus write (16) as

$$\begin{aligned}\sum_{j=1}^{K-1}\mathbf{E}[A_j]+\mathbf{E}[G_j] &\leqslant \sum_{j=1}^{K-1}(1+H_{K-1}-H_j)+2(K-1)\\ &=\sum_{j=1}^{K-1}\left(1+\sum_{i=j+1}^{K-1}\frac{1}{i}\right)+2(K-1)\\ &=\sum_{j=1}^{K-1}(j-1)\frac{1}{j}+3(K-1)=\mathcal{O}(K). \qquad \square\end{aligned}$$

Proof of Lemma 3. The proof follows immediately from combining Lemma 5 and Lemma 9. □

10. Lower bounds

We now show that the bound in Theorem 1 is information theoretically optimal up to constant factors. The proof is similar to the lower bound proof for the standard stochastic multi-armed bandit problem. However, since we make a number of assumptions not present in the standard case (such as a total ordering of $\mathcal{B}$), we present a simple self-contained lower bound argument, rather than a reduction from the standard case.

Theorem 2. *For any fixed $\epsilon > 0$ and any algorithm ϕ for the dueling bandits problem, there exists a problem instance such that*

$$R_T^{\phi}=\Omega\left(\frac{K}{\epsilon}\log T\right),$$

where $\epsilon=\min_{b\neq b^*}P(b^*>b)$.

Here is a heuristic explanation of why we might suspect the theorem to be true. Rather than consider the general problem of identifying the best of K bandits, suppose we are given a bandit b, and asked to determine with probability at least $1-1/T$ whether b is the best bandit. (Intuitively, the regret incurred by the optimal algorithm for this decision problem should be a lower bound on the regret incurred by the optimal algorithm for the general problem.) We have seen that, given two bandits b_i and b_j with $P(b_i>b_j)=1/2+\epsilon$, we can identify the better bandit with probability at least $1-1/T$ after $O(\log T/\epsilon^2)$ comparisons. If this is in fact the minimum number of comparisons required, then we would suspect that any algorithm for the above decision problem that is uniformly good over all problem instances must perform $\Omega(\log T/\epsilon^2)$ comparisons involving each inferior bandit. We will see in Lemma 10 that this is in fact the case, and we begin by constructing the appropriate problem instance.

Fix $\epsilon>0$ and define the following family of problem instances. In instance j, let b_j be the best bandit, and order the remaining bandits by their indices. That is, in instance j, we have $b_j \succ b_k$ for all $k \neq j$, and for $i,k\neq j$, we have $b_i \succ b_k$ whenever $i<k$. Given this ordering, define the winning probabilities by $P(b_i>b_k)=1/2+\epsilon$ whenever $b_i\succ b_k$. Note that this construction yields a valid problem instance, i.e. one that satisfies (3), (4).

Let q_j be the distribution on T-step histories induced by a given algorithm ϕ under instance j, and let $n_{j,T}$ be the number of comparisons involving bandit b_j scheduled by ϕ up to time T. Using these instances, we prove Lemma 10, from which Theorem 2 follows.

Lemma 10. *Let ϕ be an algorithm for the dueling bandits problem such that*

$$R_T^{\phi}=o\left(T^a\right) \tag{18}$$

for all $a>0$. Then for all j,

$$\mathbf{E}_{q_1}[n_{j,T}]=\Omega\left(\frac{\log T}{\epsilon^2}\right).$$

Lemma 10 formalizes the intuition given above, in that any algorithm whose regret is $o(T^a)$ over all problem instances must make $\Omega(\log T/\epsilon^2)$ comparisons involving each inferior bandit, in expectation. The proof is motivated by Lemma 5 of [19].

Proof of Lemma 10. Fix an algorithm ϕ satisfying assumption (18), and fix $0<a<1/2$. Define the event $\mathcal{E}_j=\{n_{j,T}<\log(T)/\epsilon^2\}$, and let $J=\{j\colon q_1(\mathcal{E}_j)<1/3\}$. For each $j\in J$, we have by Markov's inequality that

$$\mathbf{E}_{q_1}[n_{j,T}] \geqslant q_1\big(\mathcal{E}_j^c\big)\big(\log(T)/\epsilon^2\big) = \Omega\left(\frac{\log T}{\epsilon^2}\right),$$

so it remains to show that $\mathbf{E}_{q_1}[n_{j,T}] = \Omega(\log T/\epsilon^2)$ for each $j \notin J$. For any j, we know that under q_j, the algorithm ϕ incurs regret ϵ for every comparison involving a bandit $b \neq b_j$. This fact together with the assumption (18) on ϕ implies that $\mathbf{E}_{q_j}[T - n_{j,T}] = o((T^a)/\epsilon)$. Using this fact, we have by Markov's inequality that

$$\begin{aligned} q_j(\mathcal{E}_j) &= q_j\big(\big\{T - n_{j,T} > T - \log(T)/\epsilon^2\big\}\big) \\ &\leqslant \frac{\mathbf{E}_{q_j}[T - n_{j,T}]}{T - \log(T)/\epsilon^2} = o\big(T^{a-1}\big), \end{aligned}$$

where the last equality follows from simplification and the fact that ϵ is a fixed constant which does not depend on T. Choosing T sufficiently large shows that $q_j(\mathcal{E}_j) < 1/3$ for each j (and in particular, that $1 \in J$ by construction). Now by Lemma 6.3 of [18], we have that for any event $\mathcal{E}$ and distributions p, q with $p(\mathcal{E}) \geqslant 1/3$ and $q(\mathcal{E}) < 1/3$,

$$KL(p; q) \geqslant \frac{1}{3}\ln\left(\frac{1}{3q(\mathcal{E})}\right) - \frac{1}{e}.$$

For each $j \notin J$, we may apply this lemma with q_1, q_j, and the event $\mathcal{E}_j$, to show

$$\begin{aligned} KL(q_1; q_j) &\geqslant \frac{1}{3}\ln\left(\frac{1}{3o(T^{a-1})}\right) - \frac{1}{e} \\ &= \Omega(\log T). \end{aligned} \tag{19}$$

On the other hand, by the chain rule for KL-divergence [11], we have

$$\begin{aligned} KL(q_1; q_j) &\leqslant \mathbf{E}_{q_1}[n_{j,T}]\,KL(1/2 + \epsilon; 1/2 - \epsilon) \\ &\leqslant 16\epsilon^2 \mathbf{E}_{q_1}[n_{j,T}], \end{aligned} \tag{20}$$

where we use the shorthand $KL(1/2 + \epsilon; 1/2 - \epsilon)$ to denote the KL-divergence between two Bernoulli distributions with parameters $1/2 + \epsilon$ and $1/2 - \epsilon$, respectively. The first inequality follows from the fact that if a comparison does not involve bandit b_j, then the distribution on the outcome of that comparison will be the same under distributions q_1 and q_j. To see this, recall that by construction, under distribution q_1, the bandits are ordered by their indices, so that for any two bandits b_i and b_k, $q_1(b_i > b_k) = 1/2 + \epsilon$ if and only if $i < k$. On the other hand, under distribution q_j, bandit b_j is the best bandit, with all other bandits ordered by their indices, and so by construction, for any $i, k \neq j$, we have $q_j(b_i > b_k) = 1/2 + \epsilon$ if and only if $i < k$. Thus, any comparison that does not involve bandit b_j will have the same distribution on its outcome under q_1 and q_j.

The second inequality follows from a standard result on the KL-divergence between two Bernoulli distributions. Combining (19) and (20) shows that $\mathbf{E}_{q_1}[n_{j,T}] = \Omega(\log T/\epsilon^2)$ for each $j \notin J$, which proves the lemma. □

Proof of Theorem 2. Let ϕ be any algorithm for the dueling bandits problem. If ϕ does not satisfy the hypothesis of Lemma 10, the theorem holds trivially. Otherwise, on the problem instance specified by q_1, ϕ incurs regret at least ϵ every time it plays a match involving $b_j \neq b_1$. It follows from Lemma 10 that

$$R_T^{\phi} \geqslant \sum_{j \neq 1} \epsilon \mathbf{E}_{q_1}[n_{j,T}] = \Omega\left(\frac{K}{\epsilon}\log T\right). \qquad \square$$

11. Discussion of related work

Algorithms for finding maximal elements in a noisy information model are discussed in [14]. That paper describes a tournament-style algorithm that returns the best of K elements with probability $1 - \delta$ in $O(K\log(1/\delta)/\epsilon^2)$ comparisons, where ϵ is the minimum margin of victory of one element over an inferior one. This is achieved by arranging the elements in a binary tree and running a series of mini-tournaments, in which a parent and its two children compete until a winner can be identified with high confidence. Winning nodes are promoted to the parent position, and lower levels of the tree are pruned to reduce the total number of comparisons. The maximal element eventually reaches the root of the tree with high probability.

Such a tournament could incur very high regret in our framework. Consider a mini-tournament involving three suboptimal but barely distinguishable elements (e.g. $P(b^* > b_{i,j,k}) \approx 1$, but $P(b_i > b_j) = 1/2 + \gamma$ for $\gamma \ll 1$). This tournament would require $\Omega(1/\gamma^2)$ comparisons to determine the best element, but each comparison would contribute $\Omega(1)$ to the total regret. Since γ can be arbitrarily small compared to $\epsilon^* = \epsilon_{1,2}$, this yields a regret bound that can be arbitrarily worse than the above lower bound. In general, algorithms that achieve low regret in our model must avoid such situations, and must discard suboptimal bandits after as few comparisons as possible. This heuristic motivates the interleaved structure of Interleaved Filter, which allows for good control over the number of matches involving suboptimal bandits.

As mentioned in Section 2, there are striking similarities between Interleaved Filter and the Successive Elimination algorithm [12] for multi-armed bandit problems in the PAC setting, and there are also some key differences. The most obvious difference is the fact that Interleaved Filter maintains an "incumbent" arm in each round that participates in many more samples than the other arms, which is motivated by the need to eliminate highly suboptimal arms quickly in the regret minimization setting. The Successive Elimination algorithm, on the other hand, samples every bandit arm uniformly in every round. It would be interesting to see if a similar algorithm could also achieve optimal regret in the dueling bandits regret minimization setting.

12. Conclusion

We have proposed a novel framework for partial information online learning in which feedback is derived from pairwise comparisons, rather than absolute measures of utility. We have defined a natural notion of regret for this problem, and designed an algorithm that is information theoretically optimal for this performance measure. Our results extend previous work on computing in noisy information models, and are motivated by practical considerations from information retrieval applications. Future directions include finding other reasonable notions of regret in this framework (e.g., via contextualization [22]), and designing algorithms that achieve low-regret when the set of bandits is very large (a special case of this is addressed in [29]).

Acknowledgments

The work is funded by NSF Awards IIS-0812091 and IIS-0905467. The first author is also supported by a Microsoft Research Graduate Fellowship and a Yahoo! Key Scientific Challenges Award. The third author is supported by NSF Awards CCF-0643934 and AF-0910940, an Alfred P. Sloan Foundation Fellowship, and a Microsoft Research New Faculty Fellowship.

Appendix A. Satisfying modeling assumptions

The following lemma describes a general family of probabilistic comparison models and proves that strong stochastic transitivity and stochastic triangle inequality are both satisfied by this family of models. Note that both the logistic and Gaussian models described in Section 3.1 are contained within this family of models.

Lemma 11. *Let each bandit $b_i \in \{b_1 \ldots b_K\}$ be associated with a distinct real value μ_i such that $b_i \succ b_j \Leftrightarrow \mu_i > \mu_j$, and that outcomes from comparing two bandits are determined by*

$$P(b_i > b_j) = \sigma(\mu_i - \mu_j),$$

for some transfer function σ. Let σ satisfy the following properties:

- *σ is monotonically increasing.*
- *$\sigma(-\infty) = 0$.*
- *$\sigma(\infty) = 1$.*
- *$\sigma(x) = 1 - \sigma(-x)$ (rotation symmetric).*
- *$\sigma(x)$ has a single inflection point at $\sigma(0) = 1/2$.*

Then these probabilistic comparisons satisfy strong stochastic transitivity and stochastic triangle inequality.

Proof. We begin by noting that these properties essentially mean that σ behaves like a symmetric cumulative distribution function with a single inflection point at $\sigma(0) = 1/2$ (i.e., σ is an "S-shaped" curve).

For any triplet of bandits $b_i \succ b_j \succ b_k$, we know that $\mu_i > \mu_j > \mu_k$. To show strong stochastic transitivity, we first note that σ is monotonically increasing. Thus we know that $\sigma(\mu_i - \mu_k) \geqslant \sigma(\mu_i - \mu_j)$ and $\sigma(\mu_i - \mu_k) \geqslant \sigma(\mu_j - \mu_k)$, which implies that

$$\begin{aligned} \epsilon_{i,k} &= \sigma(\mu_i - \mu_k) - \frac{1}{2} \\ &\geqslant \max\left\{\sigma(\mu_i - \mu_j) - \frac{1}{2}, \sigma(\mu_j - \mu_k) - \frac{1}{2}\right\} \\ &= \max\{\epsilon_{i,j}, \epsilon_{j,k}\}. \end{aligned}$$

Round r	1	2	3	4	5	6	7	8
IF incumbent $b^{(r)}$	b_{500}	b_{150}	b_{80}	b_{25}	b_{10}	b_6	b_2	b_1
Random walk incumbent $\tilde{b}^{(r)}$	b_{500}	b_{250}	b_{120}	b_{65}	b_{35}	b_{15}	b_9	b_5

Fig. 2. Showing an example coupled sequence for IF and the Random Walk Model drawn using a measure space defined in Definition 2. At each round r, the incumbent $b^{(r)}$ of IF is always superior to the incumbent $\tilde{b}^{(r)}$ of the Random Walk Model, i.e. $b^{(r)} \succsim \tilde{b}^{(r)}$.

To show stochastic triangle inequality, we first note that $\sigma(x)$ is sub-additive, or concave, for $x \geqslant 0$. Define

$$\alpha = \frac{\mu_i - \mu_j}{\mu_i - \mu_k}$$

such that $(\mu_i - \mu_j) = \alpha(\mu_i - \mu_k)$ and $(\mu_j - \mu_k) = (1-\alpha)(\mu_i - \mu_k)$. Then we know from concavity of σ that

$$\alpha\sigma(\mu_i - \mu_k) + (1-\alpha)\sigma(0) \leqslant \sigma(\mu_i - \mu_j),$$

and also

$$(1-\alpha)\sigma(\mu_i - \mu_k) + \alpha\sigma(0) \leqslant \sigma(\mu_j - \mu_k).$$

Adding the two inequalities above yields

$$\sigma(\mu_i - \mu_k) + \sigma(0) \leqslant \sigma(\mu_i - \mu_j) + \sigma(\mu_j - \mu_k),$$

and thus

$$\epsilon_{i,k} \leqslant \epsilon_{i,j} + \epsilon_{j,k}. \qquad \square$$

Appendix B. Analyzing the Random Walk Model

Let $b^{(r)}$ and $\tilde{b}^{(r)}$ denote the candidate bandits in round r of IF and the Random Walk Model described in Definition 1, respectively. Our analysis approach leverages a particular type of measure space (defined in Definition 2 below) in order to construct a stochastic coupling between IF and the Random Walk Model. This will allow us to draw coupled sequences of candidates from the execution histories of IF and the Random Walk Model simultaneously. Fig. 2 shows an example pair of sequences drawn using our measure space. Note that $b^{(r)} \succsim \tilde{b}^{(r)}$ at any round r. We will show in the following that:

- Measure spaces from Definition 2 define the correct distribution of sequences of candidates for IF.
- Measure spaces from Definition 2 define the correct distribution of sequences of candidates for the Random Walk Model.
- Measure spaces from Definition 2 can be used to draw coupled sequences (one each from IF and the Random Walk Model) such that $b^{(r)} \succsim \tilde{b}^{(r)}$ at any round r in any pair of sequences drawn.
- At least one such measure space exists.

This implies that the sequence of candidates in executions of IF is stochastically dominated by random walks in the Random Walk Model. Hence, any bound on the random walk path length in the Random Walk Model also bounds the number of rounds in IF.

Definition 2. We define a family of measure spaces $\mathcal{M}$ in the following way. Each point in the sample space is a joint realization of the sequences of random variables X_{ij}^{rt} and Z_i^r for every pair of bandits b_i and b_j, and positive integers r and t. We will define a joint distribution over the random variables X_{ij}^{rt} and a conditional distribution over the Z_i^r variables given the X_{ij}^{rt} variables. The random variables and their distributions are explained in greater detail below.

- For every pair of bandits b_i, b_j, and positive integer r, there is a sequence of Bernoulli random variables X_{ij}^{rt} (for $t = 1, 2, \ldots$) describing the outcomes of comparisons in a match played by b_i and b_j in round r provided that b_i is the incumbent in that round. In particular $X_{ij}^{rt} = 1$ if b_i wins the tth comparison between b_j in round r, and $X_{ij}^{rt} = 0$ if b_i loses that comparison (i.e. $P(X_{ij}^{rt} = 1) = 1/2 + \epsilon_{i,j}$). We will also define the following useful notation to denote prior execution histories: $\mathcal{X}_i^r$ is the σ-field generated by the random variables $\{X_{ij}^{qt}\colon\ j \neq i,\ q < r,\ t = 1, 2, \ldots\}$.
- For a fixed i, the random variables X_{ij}^{rt} are all mutually independent as one varies j, r, t, and they have the correct distribution for each pair i, j. (In other words, the probability of b_i beating b_j is $1/2 + \epsilon_{ij}$.)
- For convenience we also define Y^r, for every positive integer r, to denote the identity of the incumbent in round $r+1$ (i.e., the bandit that wins round r) when running algorithm IF with the comparison outcomes specified by $\{X_{ij}^{rt}\}$. Note that the value (likewise distribution) of Y^r is completely determined by the values (joint distribution) of X_{ij}^{rt}.

- For every bandit b_i and positive integer r, there is a random variable Z_i^r taking non-negative integer values, such that the distribution of $Y^r + Z_i^r$, conditioned on $\mathcal{X}_i^r$, is uniform on $1, \ldots, i-1$ at every sample point where $Y^{r-1} \leqslant i$ and IF does not make a mistake in rounds $1, \ldots, r$. (This will later be used to show that the Random Walk Model stochastically dominates any mistake-free execution of IF.)

The values of X_{ij}^{rt} completely determine the history of execution of IF.[6] Our independence assumptions ensure that the history of play observed by IF has the correct distribution over histories.

Again, let $b^{(r)}$ and $\tilde{b}^{(r)}$ denote the candidate bandits in round r of IF and the Random Walk Model, respectively. By construction, $b^{(r)}$ follows the distribution defined by Y^r. We can show that $\tilde{b}^{(r)}$ follows the distribution defined by $Y^r + Z_i^r$. As alluded to in the beginning of this section, this implies that $b^{(r)} \succcurlyeq \tilde{b}^{(r)}$ at any round r.

A priori, it is not obvious that measure spaces $\mathcal{M}$ satisfying Definition 2 exist; the constraint on the conditional distribution of $Y^r + Z_i^r$ is non-trivial but we prove below that it is possible to design a measure space that satisfies this constraint, i.e. $\mathcal{M}$ is not empty. We will then show how any measure space in $\mathcal{M}$ defines a stochastic coupling between the number of rounds required in mistake-free executions of IF and the length of random walks in the Random Walk Model. To begin proving that $\mathcal{M}$ is non-empty, we first prove in Lemma 12 a constraint on the distribution of the Y^r variables. This requires us to introduce the following notation.

Definition 3. Let $U(i,k,r,t \mid \mathcal{X}_i^r)$ denote the collection of comparison sequences of length t in round r between the incumbent b_i and each other remaining b_j which results in b_k being declared the winner after t comparisons. In other words, an element in $U(i,k,r,t \mid \mathcal{X}_i^r)$ consists of a realization of each $X_{ij}^{t'r}$ for incumbent b_i, all remaining b_j, and time steps $1 \leqslant t' \leqslant t$.

Lemma 12. *For any measure space in $\mathcal{M}$, we have*

$$\forall r,\ \forall j \in \{1,2,\ldots,i-1\}:\quad \sum_{j'=1}^{j} P\big(Y^r = j' \,\big|\, \mathcal{X}_i^r, N^r\big) \geqslant \frac{j}{i-1}, \tag{B.1}$$

where b_i denotes the incumbent bandit chosen by IF for round r, the Y^r and X_{ij}^{rt} variables and the $\mathcal{X}_i^r$ σ-field are defined as in Definition 2, and N^r denotes the event that IF does not make a mistake in round r.

Proof. We first define N^{rti} as the event that IF does not make a mistake in round r, that b_i is the incumbent in that round, and that IF makes exactly t comparisons between b_i and each other remaining bandit in round r. This notation will be used throughout the proof.

The proof can be decomposed into two stages. In Stage 1, we present a series of reductions and show that it suffices to prove (B.4) below. In Stage 2, we prove (B.4).

Stage 1, *reduction* 1. To prove (B.1), it suffices to prove the following inequality,

$$\forall t \geqslant t_{\min},\ \forall r,\ \forall j \in \{1,2,\ldots,i-1\}:\quad \sum_{j'=1}^{j} P\big(Y^r = j' \,\big|\, \mathcal{X}_i^r, N^{rti}\big) \geqslant \frac{j}{i-1}, \tag{B.2}$$

where $t_{\min}$ denotes the minimum number of comparisons required for IF to determine a winner. Since (B.2) will be shown to apply for all feasible t, i, then (B.1) will also hold.

Stage 1, *reduction* 2. To prove (B.2), it suffices to show that

$$\forall 1 \leqslant j < k < i:\quad P\big(Y^r = j \,\big|\, \mathcal{X}_i^r, N^{rti}\big) \geqslant P\big(Y^r = k \,\big|\, \mathcal{X}_i^r, N^{rti}\big), \tag{B.3}$$

since then (B.2) follows from iteratively applying the pigeonhole principle (for $j = 1, \ldots, i-1$), and noting that

$$\sum_{j'=1}^{i-1} P\big(Y^r = j' \,\big|\, \mathcal{X}_i^r, N^{rti}\big) = 1.$$

Stage 1, *reduction* 3. Let $U(i,k,r,t \mid \mathcal{X}_i^r)$ be defined as in Definition 3. To prove (B.3), we will define a bijection between $U(i,j,r,t \mid \mathcal{X}_i^r)$ and $U(i,k,r,t \mid \mathcal{X}_i^r)$ for $j < k$ such that

$$P\big(U\big(i,j,r,t \,\big|\, \mathcal{X}_i^r\big) \,\big|\, \mathcal{X}_i^r, N_{rt}\big) \geqslant P\big(U\big(i,k,r,t \,\big|\, \mathcal{X}_i^r\big) \,\big|\, \mathcal{X}_i^r, N^{rti}\big). \tag{B.4}$$

[6] Some of the values X_{ij}^{rt} are exposed as IF runs and schedules matches. Other values never get exposed. In particular, for pairs of bandits b_i and b_j where neither is the incumbent in round r, the values X_{ij}^{rt} have no bearing on the history of play observed by IF.

It is straightforward to see that

$$P\big(Y^r = k \mid \mathcal{X}_i^r, N^{rti}\big) = P\big(U\big(i,k,r,t \mid \mathcal{X}_i^r\big) \mid \mathcal{X}_i^r, N^{rti}\big), \tag{B.5}$$

since $U(i,k,r,t \mid \mathcal{X}_i^r)$ contains exactly the comparison sequences that result in $Y^r = k$ given incumbent b_i in round r. Combined with (B.4), this directly implies (B.3), and thus completes the proof.

Stage 2, *proving* (B.4). The bijection is defined as follows. Each $u_k \in U(i,k,r,t \mid \mathcal{X}_i^r, N^{rti})$ is mapped to the corresponding $u_j \in U(i,j,r,t \mid \mathcal{X}_i^r, N^{rti})$ that consists of the same sequence of comparison realizations as u_k, except that the comparison realizations involving b_j and b_k are swapped (implying that b_j is declared the winner).

To prove (B.4), it remains to show that $P(u_j \mid \mathcal{X}_i^r, N^{rti}) \geqslant P(u_k \mid \mathcal{X}_i^r, N^{rti})$ for all u_j, u_k pairings in the bijection. In the sequence of comparisons defined by u_k, let

$$A = \sum_{t'=1}^{t} X_{ik}^{rt'} \quad \text{and} \quad B = \sum_{t'=1}^{t} X_{ij}^{rt'}.$$

Note that $A < B$ since b_k wins the round.[7] Also note that A and B are fixed for any u_k.[8] Under the corresponding u_j, the two summations are reversed,

$$B = \sum_{t'=1}^{t} X_{ik}^{rt'} \quad \text{and} \quad A = \sum_{t'=1}^{t} X_{ij}^{rt'},$$

and all other sequences of variables $X_{ii'}^{rt'}$ for $i' \neq j$, $i' \neq k$ remain the same.

Note that $P(X_{ik}^{rt} = 1) \geqslant P(X_{ij}^{rt} = 1)$, since b_k is inferior to b_j.[9] We define $p = P(X_{ij}^{rt} = 1)$ and $q = P(X_{ik}^{rt} = 1)$. Since all the $X_{ii'}^{rt'}$ variables are mutually independent, we can write the ratio of the conditional probabilities of u_j and u_k as

$$\begin{aligned}
\frac{P(u_j \mid \mathcal{X}_i^r, N^{rti})}{P(u_k \mid \mathcal{X}_i^r, N^{rti})} &= \frac{P(\sum_{t=1}^{t'} X_{ij}^{rt} = A) P(\sum_{t=1}^{t'} X_{ik}^{rt} = B)}{P(\sum_{t=1}^{t'} X_{ij}^{rt} = B) P(\sum_{t=1}^{t'} X_{ik}^{rt} = A)} \\
&= \frac{p^A (1-p)^{t'-A} q^B (1-q)^{t'-B}}{p^B (1-p)^{t'-B} q^A (1-q)^{t'-A}} \\
&= \frac{p^{A-B} (1-q)^{A-B}}{q^{A-B} (1-p)^{A-B}} \geqslant 1,
\end{aligned}$$

where the first equality follows from noting that all comparisons are independent and canceling out common terms (i.e., the realizations of $X_{ii'}^{rt}$ for $i' \neq j$ and $i' \neq k$), and the last inequality follows from noting that $A < B$ and $p \leqslant q$. This immediately implies (B.4), and thus completes the proof. □

Corollary 2. *For the setting described in Lemma* 12, *we also have*

$$\forall r,\ \forall 1 \leqslant j < i \leqslant i': \quad \sum_{j'=1}^{j} P\big(Y^r = j' \mid \mathcal{X}_i^r, N^r\big) \geqslant \frac{j}{i'-1}.$$

Lemma 13. *The family of measure spaces $\mathcal{M}$ defined in Definition* 2 *is non-empty.*

Proof. We will use the notation for X_{jk}^{rt}, Y^r, Z_j^r, $\mathcal{X}_j^r$ as described in Definition 2. We will show that it is possible to construct a distribution on the non-negative random variables Z_j^r which satisfies the requirements of Definition 2. Since we are conditioning on X_{ij}^{qt} for all $q < r$, then the value of Y^{r-1} is fixed (i.e., we know who the incumbent is in round r). Assume WLOG that $Y^{r-1} = i$ (i.e., the incumbent in round r is b_i). We will construct Z_i^r based on the following two cases.

Case 1. IF does not make a mistake in round r and $Y^{r-1} \leqslant i$ (meaning the incumbent during round r was b_i). We will use the following flow network to construct the conditional distribution of $Y^r + Z_i^r$ (given $\mathcal{X}_i^r$ and N^r),

- source s and sink t,

[7] The incumbent b_i wins the least number of comparisons with b_k versus any other bandit.

[8] In Definition 3, each element u_k consists of a realization of all the comparisons. Thus A and B are fixed for any u_k.

[9] By construction $P(X_{ik}^{rt} = 1) = 1/2 + \epsilon_{i,k}$.

- vertices $u_1, \ldots, u_{i-1}$,
- vertices $v_1, \ldots, v_{i-1}$,
- edges from s to each u_j with capacity $P(Y^r = j \mid \mathcal{X}_i^r, N^r)$,
- edges from each u_j to v_k where $k \geqslant j$ with infinite capacity,
- edges from each v_k to t with capacity $1/(i-1)$.

Lemma 12 and Corollary 2 imply that the minimum s–t cut of this network has capacity 1, and consequently the maximum s–t flow has value 1. In any maximum flow, each edge (s, u_j) and each edge (v_j, t) (for $1 \leqslant j \leqslant i-1$) must be saturated. Given a maximum flow, we can interpret the flow on the edge from u_j to v_k to be the joint conditional probability $P(Y^r = j, Z_i^r = k - j \mid \mathcal{X}_i^r, N^r)$, from which we can recover the conditional distribution of Z_i^r given $\mathcal{X}_i^r$ and N^r. The fact that the conditional distribution of $Y^r + Z_i^r$ is uniform on $1, \ldots, i-1$, given $\mathcal{X}_i^r, N^r$, follows from the fact that the flow from v_k to t is exactly $1/(i-1)$ for every k.

Case 2. IF does make a mistake in round r or $Y^{r-1} > i$. Then we set Z_i^r to some arbitrary non-negative integer, e.g., 0.

Thus, we have shown that there exists a feasible probability distribution on the Z_i^r variables which satisfies the requirements of Definition 2. This implies that $\mathcal{M}$ is non-empty. □

Lemma 14. *The number of rounds in mistake-free executions of IF is stochastically dominated by the length of random walks in the Random Walk Model.*

Proof. Let $b^{(r)}$ and $\tilde{b}^{(r)}$ be the incumbents chosen by IF and the Random Walk Model, respectively, at round r. It suffices to construct a stochastic coupling between IF and the Random Walk Model such that $b^{(r)} \succcurlyeq \tilde{b}^{(r)}$ for all r.

We can take any measure space in $\mathcal{M}$ to construct this stochastic coupling, and we know from Lemma 13 that at least one such measure space exists. There is one sample point for every possible joint outcome of the random variables X_{ij}^{rt} and Z_i^r. The execution of IF is determined by the X_{ij}^{rt} variables (note that $b^{(r)} = b_{i'}$ where $i' = Y^{r-1}$). Consider any execution of IF that is mistake-free through rounds $1, \ldots, s$. The analogous execution of the Random Walk Model is determined by looking at the sequence of incumbents when one runs a “perturbed” version of IF. The perturbation consists to taking the identity of the incumbent in round $r+1$ (for every $r = 1, \ldots, s$) and modifying it by adding Z_i^r (where b_i is the incumbent of “perturbed” IF in round r), and then executing round $r+1$ using the perturbed incumbent instead of the one that would ordinarily be chosen by IF. Both IF and “perturbed” IF start with the same initial incumbent at the beginning of round 1 chosen uniformly from $1, \ldots, K$.

It is now straightforward to see that this stochastic coupling holds from the definition of the Y^r and Z_i^r variables in Definition 2, so long as the initial condition $b^{(1)} \succcurlyeq \tilde{b}^{(1)}$ holds. We finally note that $b^{(1)} = \tilde{b}^{(1)}$ by construction, which proves the theorem. □

References

[1] Jean-Yves Audibert, Sébastien Bubeck, Minimax policies for adversarial and stochastic bandits, in: Conference on Learning Theory (COLT), 2009.

[2] Peter Auer, Nicolò Cesa-Bianchi, Paul Fischer, Finite-time analysis of the multiarmed bandit problem, Mach. Learn. 47 (2) (2002) 235–256.

[3] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, Robert Schapire, The nonstochastic multiarmed bandit problem, SIAM J. Comput. 32 (1) (2002) 48–77.

[4] Micah Adler, Peter Gemmell, Mor Harchol-Balter, Richard Karp, Claire Kenyon, Selection in the presence of noise: The design of playoff systems, in: ACM–SIAM Symposium on Discrete Algorithms (SODA), 1994.

[5] Nir Ailon, Mehryar Mohri, An efficient reduction of ranking to classification, in: Conference on Learning Theory (COLT), 2008.

[6] Peter Auer, Using confidence bounds for exploitation-exploration trade, J. Mach. Learn. Res. 3 (2003) 397–422.

[7] Maria-Florina Balcan, Nikhil Bansal, Alina Beygelzimer, Don Coppersmith, John Langford, Gregory Sorkin, Robust reductions from ranking to classification, in: Conference on Learning Theory (COLT), 2007.

[8] Michael Ben-Or, Avinatan Hassidim, The bayesian learner is optimal for noisy binary search (and pretty good for quantum as well), in: IEEE Symposium on Foundations of Computer Science (FOCS), 2008.

[9] Nicolò Cesa-Bianchi, Gábor Lugosi, Gilles Stoltz, Regret minimization under partial monitoring, Math. Oper. Res. 31 (3) (2006) 562–580.

[10] William Cohen, Robert Schapire, Yoram Singer, Learning to order things, J. Artificial Intelligence Res. 10 (1999) 243–270.

[11] Thomas M. Cover, Joy A. Thomas, Elements of Information Theory, J. Wiley, 1999.

[12] Eyal Even-Dar, Shie Mannor, Yishay Mansour, Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems, J. Mach. Learn. Res. 7 (2006) 1079–1105.

[13] Yoav Freund, Raj Iyer, Robert Schapire, Yoram Singer, An efficient boosting algorithm for combining preferences, J. Mach. Learn. Res. 4 (2003) 933–969.

[14] Uriel Feige, Prabhakar Raghavan, David Peleg, Eli Upfal, Computing with noisy information, SIAM J. Comput. 23 (5) (1994).

[15] Ralf Herbrich, Thore Graepel, Klaus Obermayer, Support vector learning for ordinal regression, in: International Conference on Artificial Neural Networks (ICANN), 1999.

[16] Wassily Hoeffding, Probability inequalities for sums of bounded random variables, J. Amer. Statist. Assoc. 58 (1963) 13–30.

[17] Thorsten Joachims, A support vector method for multivariate performance measures, in: International Conference on Machine Learning (ICML), 2005.

[18] Richard M. Karp, Robert Kleinberg, Noisy binary search and its applications, in: ACM–SIAM Symposium on Discrete Algorithms (SODA), 2007.

[19] Robert Kleinberg, Alexandru Niculescu-Mizil, Yogeshwer Sharma, Regret bounds for sleeping experts and bandits, in: Conference on Learning Theory (COLT), 2008.

[20] T.L. Lai, Herbert Robbins, Asymptotically efficient adaptive allocation rules, Adv. in Appl. Math. 6 (1985) 4–22.

[21] Phil Long, Rocco Servedio, Boosting the area under the ROC curve, in: Proceedings of Neural Information Processing Systems (NIPS), 2007.

[22] John Langford, Tong Zhang, The epoch-greedy algorithm for contextual multi-armed bandits, in: Proceedings of Neural Information Processing Systems (NIPS), 2007.

[23] Rajeev Motwani, Prabhakar Raghavan, Randomized Algorithms, Cambridge University Press, 1995.

[24] Shie Mannor, John N. Tsitsiklis, The sample complexity of exploration in the multi-armed bandit problem, J. Mach. Learn. Res. 5 (2004) 623–648.

[25] Sandeep Pandey, Deepak Agarwal, Deepayan Chakrabarti, Vanja Josifovski, Bandits for taxonomies: A model-based approach, in: SIAM Conference on Data Mining (SDM), 2007.

[26] Filip Radlinski, Madhu Kurup, Thorsten Joachims, How does clickthrough data reflect retrieval quality?, in: ACM Conference on Information and Knowledge Management (CIKM), 2008.

[27] Herbert Robbins, Some aspects of the sequential design of experiments, Bull. Amer. Math. Soc. 58 (1952) 527–535.

[28] Yisong Yue, Josef Broder, Robert Kleinberg, Thorsten Joachims, The k-armed dueling bandits problem, in: Conference on Learning Theory (COLT), 2009.

[29] Yisong Yue, Thorsten Joachims, Interactively optimizing information retrieval systems as a dueling bandits problem, in: International Conference on Machine Learning (ICML), 2009.